Distributional regimes for the number of *k*-word matches between two random sequences

Ross A. Lippert[†‡], Haiyan Huang[§], and Michael S. Waterman[†¶]

[†]Informatics Research, Celera Genomics, Rockville, MD 20878; [§]Department of Biostatistics, Harvard University, Boston, MA 02115; and [¶]Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089-1113

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected on May 1, 2001.

Contributed by Michael S. Waterman, August 5, 2002

When comparing two sequences, a natural approach is to count the number of *k*-letter words the two sequences have in common. No positional information is used in the count, but it has the virtue that the comparison time is linear with sequence length. For this reason this statistic D_2 and certain transformations of D_2 are used for EST sequence database searches. In this paper we begin the rigorous study of the statistical distribution of D_2. Using an independence model of DNA sequences, we derive limiting distributions by means of the Stein and Chen–Stein methods and identify three asymptotic regimes, including compound Poisson and normal. The compound Poisson distribution arises when the word size *k* is large and word matches are rare. The normal distribution arises when the word size is small and matches are common. Explicit expressions for what is meant by large and small word sizes are given in the paper. However, when word size is small and the letters are uniformly distributed, the anticipated limiting normal distribution does not always occur. In this situation the uniform distribution provides the exception to other letter distributions. Therefore a naive, one distribution fits all, approach to D_2 statistics could easily create serious errors in estimating significance.

1. Introduction

Sequence comparison and database searching are among of the most frequent and useful activities in computational biology and bioinformatics. The goal is to discover relationships between sequences and thus to suggest biological features previously unknown. These searches are based on the local alignment or Smith–Waterman algorithm (1), which in important heuristic versions has been utilized in the very popular search algorithms FASTA and BLAST (2–4).

As the sizes of biological sequence databases grow, even more efficient comparison methods are required to carry out the large number of comparisons. There are effectively two components to all these methods: the discovery of approximately matching words in the sequences and the evaluation of the statistical significance of the found matching. These estimates of statistical significance can be based on Poisson approximation (ref. 5, Chap. 11). Another method of comparison is by sequence composition, using statistical features of the sequences to determine relationships. The correlation between the occurrence of various words such as AAA and AAT makes this a challenging problem. In ref. 5, Chap. 12, the reader can find the limiting multivariate normal distribution for a collection of words when the sequence length becomes large. The use of these techniques for comparisons has been somewhat limited. However, one such measure proposed originally in ref. 6, called D_2 in ref. 7, is based on the pairwise compositional (dis)similarity of two sequences, and this method has found wide application, especially for EST databases. A high-performance implementation of D_2 is at the core of the EST clustering strategy of the STACK human gene index (8, 9).

D_2 is based on a comparison of the counts of identical k-words in a pair of sequences, without producing an alignment, and it can be computed in linear time. Each sequence is associated with a vector of its k-word counts, and the Euclidean distance between these vectors provides the comparison. We define the D_2 statistic to be the number of k-word matches between the two sequences without reference to the order of occurrence of the k-words. D_2 was experimentally studied in ref. 10 and a windowed version is used as the basis for alternative splicing clustering in refs. 11 and 12. Alternative quadratic forms and metrics, related to the multivariate distribution cited above, including the Mahalanobis distance, have been experimentally explored in refs. 13 and 14. With ref. 15, we see the introduction of alternative sequence models as well.

Although the computational results on biological sequences for D_2 are numerous, with ever more sophisticated models and comparisons, until now there has been no rigorous statistical analysis of D_2 and its relatives in terms of its limiting distribution as a random variable. It is classical in statistics that the binomial distribution with n trials and success probability p has two asymptotic limiting distributions, Poisson if the product of the np tends to a positive finite limit (p must be a function of n) and normal when p does not change with n. However, as seen in *Section 5*, D_2 in the case of a two-letter alphabet and $k = 1$ has neither a Poisson nor a normal limit.

In this paper we begin the rigorous study of D_2. Admittedly, we do not give a complete analysis of the more recent D_2 variants. Our sequence model will be independent letters, and our random variable will be the inner product of the two count vectors. For this model, we derive limiting distributions by means of the Stein and Chen–Stein methods and identify three asymptotic regimes, including Poisson and normal. Stein's work allows us to give bounds on the distributional approximations. Additionally, numerical results will be supplied, which suggest that substantial improvements on our theoretical bounds are possible.

We have confined our results to those relevant to nucleotide sequence alphabets (four-letter alphabet) because this is where the important biological applications will be. The generalizations to other alphabets are not difficult to make.

2. Some Preliminaries

C. Stein introduced a revolutionary method to prove normal approximations to sums of random variables in ref. 16, and later his student L. Chen extended the concepts to Poisson approximation (17). In our setting it is possible to show both methods

[‡]To whom correspondence should be addressed. E-mail: ross.lippert@celera.com.

can be applied (for different k of course). In this section we will derive and cite some calculations and bounds that will be useful in what follows. Most of the notation and definitions used below are from ref. 5, Chap. 11.

For the two sequences with i.i.d. (independent and identically distributed) letters $\mathbf{A} = A_1A_2 \cdots A_n$ and $\mathbf{B} = B_1B_2 \cdots B_m$ of letters from finite alphabet $\mathcal{A}$, let

$$f_a = P(A_i = \mathrm{a}) = P(B_j = \mathrm{a}), \qquad a \in \mathcal{A}$$

and

$$p_k = \sum_{a \in \mathcal{A}} f_a^k .$$

Further, define the match indicator $C_{i,j} = 1\{A_i = B_j\}$, and the k-word match indicator at position (i, j)

$$Y_{i,j} = C_{i,j}C_{i+1,j+1} \cdots C_{i+k-1,j+k-1} .$$

Note: $\mathbf{E}C_{i,j} = p_2$ and $\mathbf{E}Y_{i,j} = p_2^k$.

It will be convenient to let $\bar{n} = n - k + 1$ and $\bar{m} - k + 1$ when k is understood, as the terms arise frequently.

Let the index set for k-word matches be $I = \{(i, j) : 1 \le i \le \bar{n}, 1 \le j \le \bar{m}\}$. The neighborhood of dependence for index $v = (i, j) \in I$ is defined as

$$J_v = \{u = (i', j') \in I : |i - i'| \le k \text{ or } |j - j'| \le k\}.$$

It is clear that Y_u and Y_v are independent when $u \notin J_v$.

It will be useful to further subdivide the dependency structure into two sets, $J_v = J_v^a \cup J_v^c$, where J_v^c is described by $|i - i'| \le k$ and $|j - j'| > k$ or $|i - i'| > k$ and $|j - j'| \le k$ (the *crabgrass* case in ref. 5, and J_v^a is described by $|i - i'| \le k$ and $|j - j'| \le k$ (the *accordion* case).

In this framework, the D_2 statistic is the number of k-word matches between the two sequences $\mathbf{A}$ and $\mathbf{B}$ is

$$D_2 = \sum_{v \in I} Y_v .$$

Clearly,

$$\mathbf{E}D_2 = \sum_{v \in I} \mathbf{E}Y_v = \bar{n}\bar{m}p_2^k .$$

In the remainder of this section we derive bounds relevant to $\mathrm{Var}(D_2)$. Since $\mathbf{E}(Y_vY_u) = \mathbf{E}Y_v\mathbf{E}Y_u$ when $u \notin J_v$,

$$\begin{aligned}
\mathrm{Var}(D_2) &= \sum_{v \in I} \sum_{u \in I} (\mathbf{E}(Y_vY_u) - \mathbf{E}Y_v\mathbf{E}Y_u) \\
&= \sum_{v \in I} \sum_{u \in J_v} (\mathbf{E}(Y_vY_u) - \mathbf{E}Y_v\mathbf{E}Y_u) \\
&= \sum_{v \in I} \sum_{u \in J_v} (\mathbf{E}(Y_vY_u) - p_2^{2k}).
\end{aligned}$$

Thus, for the variance of D_2, we may focus on the calculation of $\mathbf{E}(Y_vY_u)$.

When $u \in J_v^c$ we have (from ref. 5) the upper bound

$$\mathbf{E}(Y_vY_u) \le p_2^{3k/2}p_2^{3\delta k} ,$$

where $\delta \in (0, \frac{1}{6}]$, and the lower bound

$$\mathbf{E}(Y_vY_u) \ge p_2^{2k}\frac{p_3}{p_2^2} .$$

When $u \in J_v^a$, and $i - i' \ne j - j'$ we have

$$\mathbf{E}(Y_vY_u) \le p_2^{\gamma(2k+1)}p_2^{k+1/2} ,$$

where $\gamma \in (0, \frac{1}{2} - 1/(2k + 1)]$, with a lower bound given by

$$\mathbf{E}(Y_vY_u) \ge p_2^{2k} .$$

When $u \in J_v^a$, and $i - i' = j - j' = t$, we have

$$\mathbf{E}(Y_vY_u) = p_2^{k+|t|} ,$$

which is a case overlooked in ref. 5. Then $p_2^{2k} \le \mathbf{E}(Y_vY_u) \le p_2^k$. Summing $\mathbf{E}(Y_vY_u)$ over all $u \in J_v$ with $t = i - i' = j - j'$ and assuming $k > 1$, we have

$$\sum_u \mathbf{E}(Y_v Y_u) = \sum_{t=-k+1}^{k-1} p_2^{k+|t|} = p_2^k\left(2\,\frac{1-p_2^k}{1-p_2} - 1\right) \geq p_2^k\left(\frac{1+p_2-2p_2^2}{1-p_2}\right),$$

with upper bound

$$\sum_u \mathbf{E}(Y_v Y_u) \leq p_2^k\left(\frac{1+p_2}{1-p_2}\right).$$

Combining these contributions, we obtain an upper bound,

$$\sum_{v\in I}\sum_{u\in J_v} \mathbf{E}(Y_v Y_u) \leq \bar{n}\bar{m}\left\{(2k-1)(\bar{n}+\bar{m}-4k+2)p_2^{3/2k}p_2^{3\delta k} + (2k-1)(2k-2)p_2^{k+1/2}p_2^{\gamma(2k+1)} + p_2^k\frac{1+p_2}{1-p_2}\right\}, \quad [1]$$

and the following upper bound on the variance,

$$\begin{aligned}\mathrm{Var}(D_2) &= \sum_{v\in I}\sum_{u\in J_v}(\mathbf{E}(Y_v Y_u) - \mathbf{E}Y_v\mathbf{E}Y_u)\\ &\leq \bar{n}\bar{m}\left\{(2k-1)(\bar{n}+\bar{m}-4k+2)p_2^k(p_2^{1/2k} - p_2^{3\delta k} - p_2^k)\right.\\ &\quad \left. + (2k-1)(2k-2)p_2^k(p_2^{1/2}p_2^{\gamma(2k+1)} - p_2^k) + p_2^k\left(\frac{1+p_2}{1-p_2} - (2k-1)p_2^k\right)\right\}. \quad [2]\end{aligned}$$

Similarly, we can combine lower bound contributions to obtain

$$\mathrm{Var}(D_2) \geq \bar{n}\bar{m}\left\{(2k-1)(\bar{n}+\bar{m}-4k+2)p_2^{2k}(p_3/p_2^2 - 1) + p_2^k\left(\frac{1+p_2-2p_2^2}{1-p_2} - (2k-1)p_2^k\right)\right\}. \quad [3]$$

3. Poisson When $k > 2\log(n)$

When k is large enough, arguments similar to that in refs. 18 and 19 show that there are approximately a Poisson number of matching clumps. Each clump has a number of matching k-words that have a geometric distribution. Therefore the number of k-word matches is approximately a Poisson number of independent geometric random variables and such a sum is called a compound Poisson distribution. In ref. 20 there is a full treatment of the compound Poisson and sequence matching. Here are some details.

Let X_v be the declumped matching indicators associated with the Y_v by

$$\begin{aligned} X_{i,j} &= Y_{i,j} &&: \quad i = 0 \text{ or } j = 0\\ X_{i,j} &= (1 - C_{i,j})Y_{i,j} &&: \quad \text{else}\end{aligned}$$

Define a random variable D_{2*}, the "declumped D_2," as $\Sigma_{v\in I} X_v$. D_{2*} counts the number of maximal exact matches between **A** and **B** (those exact matches that cannot be extended) with length larger than k.

We can compute bounds to a Poisson approximation of D_{2*} by using the Chen–Stein theorem.

Theorem 3.1. *Let X_i for $i \in I$ be indicator random variables such that X_i is independent of $\{X_j\}$, $j \notin J_i$. Let $W = \Sigma_{i\in I} X_i$ and $\lambda = \mathbf{E}W$ and let Z be a Poisson random variable with $\mathbf{E}Z = \lambda$. Then*

$$\|W - Z\| \leq 2(b_1 + b_2)\,\frac{1-e^{-\lambda}}{\lambda} \leq 2(b_1 + b_2),$$

and in particular

$$|P(W = 0) - e^{-\lambda}| \leq (b_1 + b_2)\,\frac{1-e^{-\lambda}}{\lambda},$$

where

$$b_1 = \sum_{v\in I}\sum_{u\in J_v} \mathbf{E}X_v\mathbf{E}X_u\,,$$

and

$$b_2 = \sum_{v\in I}\sum_{v\neq u\in J_v} \mathbf{E}(X_u X_v).$$

The intensity of this Poisson process is given by the expectation of D_{2*},

$$\lambda = \mathbf{E}D_{2*} = p_2^k\{(1-p_2)\bar{n}\bar{m} + p_2(\bar{n}+\bar{m}-1)\}.$$

We obtain the upper bounds on the Chen–Stein terms for D_{2*},

$$\begin{aligned} b_1 &= \sum_{v\in I}\sum_{u\in J_v} \mathbf{E}X_v\mathbf{E}X_u \\ &\le \sum_{v\in I} \mathbf{E}X_v(2k-1)(\bar{n}+\bar{m}-2k+1)p_2^k \\ &= \lambda(2k-1)(\bar{n}+\bar{m}-2k+1)p_2^k\,. \end{aligned}$$

Because of declumping $\mathbf{E}(X_uX_v) = 0$ for $i - i' = j - j'$, $u \in J_v$. Otherwise, we take $\mathbf{E}(X_uX_v) \le \mathbf{E}(Y_uY_v)$ to make an upper bound on b_2 similar to Eq. **2**,

$$\begin{aligned} b_2 &= \sum_{v\in I}\sum_{v\ne u\in J_v} \mathbf{E}(X_uX_v) \\ &\le \sum_{v\in I}\sum_{u\in J_v} \mathbf{E}(Y_uY_v) \\ &\le \bar{n}\bar{m}(2k-1)(\bar{n}+\bar{m}-4k+2)p_2^{3/2k}p_2^{3\delta k} + (2k-1)(2k-2)p_2^{k+1/2}p_2^{\gamma(2k+1)}. \end{aligned}$$

To simplify, let $n = m$, and $k = -(\alpha/\log(p_2))\log(n)$. Additionally, we will assume $n \gg k \gg 1$ where appropriate. The parameter in Poisson approximation can be written as

$$\lambda = (1-p_2)n^{2-\alpha} = O(n^{2-\alpha}).$$

We may bound $b_1 + b_2$ by

$$b_1 + b_2 \le 4kn^{3-2\alpha} + 4kn^{3-(3/2+3\delta)\alpha} + (2k)^2p_2^{1/2+\gamma}n^{2-(2\gamma+1)\alpha}\,.$$

Thus, $b_1 + b_2$ has a rate

$$O\left(\frac{\log(n)}{n^{2\alpha-3}}\right) + O\left(\frac{\log(n)}{n^{(3/2+3\delta)\alpha-3}}\right) + O\left(\frac{(\log(n))^2}{n^{(2\gamma+1)\alpha-2}}\right).$$

In nonuniform case, clearly, $b_1 + b_2 \to 0$ when $\alpha \ge 2$. Additionally, when one chooses $\alpha = 2$, λ approaches a nonzero constant. In uniform case, since $\delta = \frac{1}{6}$ and $\gamma = \frac{1}{2} - 1/2(k+1)$, $b_1 + b_2 \to 0$ when we take $\alpha > \frac{3}{2}$. Thus we may approximate D_{2*} with a Poisson variable in a bigger regime in the uniform case.

To obtain an approximate distribution of D_2 based on D_{2*}, we assume that the k-word matches occur in isolated, independent islands. The number of the islands is approximately Poisson. The lengths of these islands are geometrically distributed according to a random variable T

$$P(\text{length} = k + T = k + t) = (1-p_2)p_2^t\,,$$

corresponding to the size of the extension past the first matching k-word.

The resulting model of D_2 is a compound Poisson process of independent geometrically distributed variables T_i,

$$D_2 \sim \sum_{i=1}^{Z(\lambda)} 1 + T_i\,,$$

where $\lambda = (1-p_2)p_2^kn^2$ and $\mathbf{E}T_i = (p_2/(1-p_2))$. See ref. 20 for a rigorous treatment of compound Poisson in this setting.

4. Normal When $k < \frac{1}{6}\log(n)$

Stein's method (16, 21) is one of the many well-known techniques for studying normal approximations. Since its introduction it has been the basis of much research and many applications. Recently, based on a differential equation and coupling, it has been applied to obtain the bounds on the distance from normality/multinormality for the sum of local dependent variables (22–24). Chen and Shao are currently working towards improving these bounds by using concentration inequalities (25).

We employ a version of Stein's method, Theorem 2.2 from ref. 22 provided below, to obtain the error bounds between the standardized D_2 and Normal.

Theorem 4.1. *Let $Y_1, \ldots, Y_n$ be random variables satisfying $|Y_i - \mathbf{E}(Y_i)| \le B$ almost surely, $i = 1, \ldots, n$, $\mathbf{E}\sum_{i=1}^n Y_i = \lambda$, $Var\sum_{i=1}^n Y_i = \sigma^2 > 0$ and $\frac{1}{n}\mathbf{E}\sum_{i=1}^n |Y_i - \mathbf{E}Y_i| = \mu$. Let $M_i \subset \{1, \ldots, n\}$ be such that $j \in M_i$ if and only if $i \in M_j$ and (Y_i, Y_j) is independent of $\{Y_k\}_{k\notin Mi\cup M_j}$ for $i, j = 1, \ldots, n$, and set $D = max_{1\le i\le n}|M_i|$. Then*

$$\left| P\left(\frac{\sum_{i=1}^n Y_i - \lambda}{\sigma} \le w\right) - \Phi(w)\right| \le 7\,\frac{n\mu}{\sigma^3}\,(DB)^2\,.$$

To obtain a normal approximation for D_2, we define a standardized auxiliary variable,

$$W = \frac{D_2 - \mathbf{E}\mathbf{D}_2}{\sqrt{\mathrm{Var}(D_2)}} = \sum_v \frac{Y_v - \mathbf{E}\mathbf{Y}_v}{\sqrt{\mathrm{Var}(D_2)}}.$$

Let $M_v = J_v$; it is easy to check that $u \in M_v$ if and only if $v \in M_u$ and (Y_u, Y_v) is independent of $\{Y_\alpha\}_{\alpha \notin M_u \cup M_v}$ for $u, v \in I$. Then for the D in Theorem *4.1*, we have $D = (2k-1)(\bar{n} + \bar{m} - 2k + 1)$. And since $|Y_u - E(Y_u)| \le 1$, $B = 1$, and $\mu \le 1$. Further, from Eq. **3**, we have

$$\begin{aligned}\sigma &= \sqrt{\mathrm{Var}(D_2)} \\ &\ge \left(\bar{n}\bar{m}\left\{(2k-1)(\bar{n}+\bar{m}-4k+2)(p_3/p_2^2 - 1)p_2^{2k} + \left(\frac{1+p_2-2p_2^2}{1-p_2} - (2k-1)p_2^k\right)p_2^k\right\}\right)^{\frac{1}{2}}.\end{aligned}$$

Substituting into *Theorem 4.1* (note that the n in the statement of the theorem becomes n^2), and letting $\bar{n} = \bar{m}$, $k = \alpha \log_{1/p_2}(n)$, we obtain a bound for the nonuniform case

$$|P(W \le w) - \Phi(w)| \le \frac{7n^2(2k-1)^2(2\bar{n}-2k+1)^2}{\left\{\bar{n}^2(2k-1)(2\bar{n}-4k+2)\left(\frac{p_3}{p_2^2}-1\right)p_2^{2k} + \bar{n}^2\left(\frac{1+p_2-2p_2^2}{1-p_2} - (2k-1)p_2^k\right)p_2^k\right\}^{\frac{3}{2}}},$$

which has a rate $O(\sqrt{\log(n)}/n^{1/2-3\alpha})$. When $\alpha < \frac{1}{6}$, the error bound is approximately zero. Thus for $k = \alpha \log_{1/p_2}(n)$ with $0 < \alpha < \frac{1}{6}$, W is approximately normal.

We arrive at our result.

Theorem 4.2. *For nonuniform i.i.d. sequences, and for $k < \frac{1}{6} \log_{1/p_2}(n)$, the D_2 statistic on k-word of sequences of length n is approximately normal.*

When the underlying sequence is uniformly distributed, $p_3/p_2^2 - 1 = 0$, then, we have

$$\sqrt{\mathrm{Var}(D_2)} \ge \left(\bar{n}^2\left(\frac{1+p_2-2p_2^2}{1-p_2} - (2k-1)p_2^k\right)p_2^k\right)^{\frac{3}{2}},$$

and we derive the rate,

$$|P(W \le w) - \Phi(w)| \le \frac{7n^2(2k-1)^2(2\bar{n}-2k+1)^2}{\left\{\bar{n}^2\left(\frac{1+p_2-2p_2^2}{1-p_2} - (2k-1)p_2^k\right)p_2^k\right\}^{\frac{3}{2}}},$$

which is asymptotically $O(\log(n)^2 n^{1+3/2\alpha})$, providing us with no bound. We therefore have proven an approach to normality only in the nonuniform case.

5. The Nonnormal Case

In numerical results shown later, D_2 has nonnormal behavior in the case of uniformly distributed letters and small k. To see just how this happens, let us consider the simplest case of $k = 1$, a binary alphabet, and the two sequences of the same length n.

Assume the alphabet is $\{0, 1\}$, and $P(0 \text{ appears}) = p$, $P(1 \text{ appears}) = q$.

Denoting the number of occurrences of 0 and 1 in the two sequences by X and Y, respectively, then

$$D_2 = XY + (n - X)(n - Y).$$

Obviously X and Y are independent binomial distributions with expectation np and variance npq. So

$$\mathbf{E}(D_2) = n^2p^2 + n^2q^2 = n^2((p+q)^2 - 2pq) = n^2(1-2pq),$$

and

$$\begin{aligned}\mathrm{Var}(D_2) &= 2n^2pq(1-2pq) + 2n^2(n-1)pq(p-q)^2 \\ &\sim \frac{n^2}{4} : p = q = \tfrac{1}{2} \\ &\sim 2pq(p-q)^2n^3 : p \ne q.\end{aligned}$$

Letting $\sigma^2 = \mathrm{Var}(D_2)$, we now consider the standardized D_2.

$$\frac{D_2 - \mathbf{E}(D_2)}{\sigma} = \frac{XY}{\sigma} + \frac{(n-X)(n-Y)}{\sigma} - \frac{n^2(1-2pq)}{\sigma}$$

$$= \frac{2XY}{\sigma} - \frac{nX}{\sigma} - \frac{nY}{\sigma} + \frac{2n^2pq}{\sigma}$$

$$= \frac{2(X-np)(Y-np)}{\sigma} + \frac{n(2p-1)(Y-np)}{\sigma} + \frac{n(2p-1)(X-np)}{\sigma}$$

$$= \frac{2npq}{\sigma}\left(\frac{X-np}{\sqrt{npq}}\right)\left(\frac{Y-np}{\sqrt{npq}}\right) + n(2p-1)\frac{\sqrt{npq}}{\sigma}\left(\frac{Y-np}{\sqrt{npq}}\right) + n(2p-1)\frac{\sqrt{npq}}{\sigma}\left(\frac{X-np}{\sqrt{npq}}\right)$$

with

$$\frac{D_2 - \mathbf{E}(D_2)}{\sigma} = \frac{2npq(X-np)}{\sigma\sqrt{npq}}\frac{(Y-np)}{\sqrt{npq}}$$

for $p = q = \frac{1}{2}$.

Note that in the uniform case,

$$\lim_{n\to\infty}\frac{2npq}{\sigma} = \lim_{n\to\infty}\frac{n/2}{n/2} = 1,$$

and $(X-np)/\sqrt{npq}$ and $(Y-np)/\sqrt{npq}$ are approximately independent $N(0,1)$ by the central limit theorem. So, the limiting distribution of $(D_2 - \mathbf{E}(D_2))/\sigma$ is $N(0,1)\cdot N(0,1)$, which obviously is not normal. In fact, the density function of the product of two independent standard normal is a Bessel function ($K_0(|x|)$).

In the nonuniform case, we have

$$\lim_{n\to\infty}\frac{npq}{\sigma} = \lim_{n\to\infty}\frac{npq}{\sqrt{2pq(p-q)^2n^3}} = 0,$$

$$\lim_{n\to\infty}\frac{n(2p-1)\sqrt{npq}}{\sigma} = \lim_{n\to\infty}\frac{n(2p-1)\sqrt{npq}}{\sqrt{2pq(p-q)^2n^3}} = \frac{2p-1}{\sqrt{2}|p-q|}.$$

So

$$\lim_{n\to\infty}\frac{D_2 - \mathbf{E}(D_2)}{\sigma} = \lim_{n\to\infty}\left(\frac{n(2p-1)\sqrt{npq}}{\sigma}\frac{(Y-np)}{\sqrt{npq}} + \frac{n(2p-1)\sqrt{npq}}{\sigma}\frac{(X-np)}{\sqrt{npq}}\right)$$

$$= \frac{2p-1}{\sqrt{2}|p-q|}(N_1 + N_2),$$

where N_1 and N_2 are independent $N(0,1)$. So the standardized D_2 is approximately normal with expectation 0 and variance $((2p-1)^2/2(p-q)^2)(1+1) = 1$. Previous work in *Section 4* provides the error bound on the normal approximation, which decays at the rate $O(1/\sqrt{n})$.

In the above results, the variance of D_2 has played an important role to form the normal or nonnormal approximation when the underlying sequence is nonuniform or uniform. In fact, for any alphabet and any length of the counted word,

$$\mathrm{Var}(D_2) = An^2 + \delta n^3,$$

where $\delta > 0$ in nonuniform case and $\delta = 0$ in uniform case.

Table 1. Kolmogorov–Smirnov *p* values for nonuniform D_2 compared with normal

k/n	$2^0 \times 10^2$	$2^1 \times 10^2$	$2^2 \times 10^2$	$2^3 \times 10^2$	$2^4 \times 10^2$	$2^5 \times 10^2$	$2^6 \times 10^2$	$2^7 \times 10^2$
1	0.05862	0.00419	0.13668	0.11486	0.31036	0.09010	0.91967	0.00506
2	0.03365	0.00006	0.00061	0.66297	0.29724	0.16957	0.66064	0.68674
3	0.00000	0.01002	0.05023	0.39328	0.05444	0.05082	0.77163	0.38298
4	0.00000	0.00004	0.00039	0.14959	0.03058	0.26901	0.59183	0.93879
5	0.00000	0.00004	0.00048	0.14381	0.03832	0.04490	0.55703	0.62759
6	0.00000	0.00000	0.00022	0.00403	0.00601	0.08003	0.59902	0.32061
7	0.00000	0.00000	0.00000	0.00009	0.00475	0.56324	0.29819	0.46705
8	0.00000	0.00000	0.00000	0.00000	0.00058	0.11751	0.32351	0.17059
9	0.00000	0.00000	0.00000	0.00000	0.00000	0.00005	0.15591	0.18042
10	0.00000	0.00000	0.00000	0.00000	0.00000	0.00002	0.02962	0.11055

Table 2. Kolmogorov–Smirnov *p* values for uniform D_2 compared with normal

k/n	$2^0 \times 10^2$	$2^1 \times 10^2$	$2^2 \times 10^2$	$2^3 \times 10^2$	$2^4 \times 10^2$	$2^5 \times 10^2$	$2^6 \times 10^2$	$2^7 \times 10^2$
1	0.00000	0.00000	0.00000	0.00000	0.00010	0.00002	0.00001	0.00002
2	0.03811	0.15773	0.28724	0.47452	0.07759	0.19055	0.25803	0.00939
3	0.04802	0.15361	0.12058	0.55153	0.70760	0.22644	0.81058	0.31066
4	0.00000	0.05730	0.04796	0.81343	0.68940	0.65794	0.98177	0.69245
5	0.00000	0.00001	0.23410	0.18908	0.77291	0.10750	0.08259	0.08706
6	0.00000	0.00000	0.00144	0.07070	0.08660	0.72020	0.06702	0.45234
7	0.00000	0.00000	0.00000	0.00000	0.02782	0.69609	0.26900	0.06839
8	0.00000	0.00000	0.00000	0.00000	0.00000	0.06281	0.65713	0.05397
9	0.00000	0.00000	0.00000	0.00000	0.00000	0.00001	0.00000	0.32139
10	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00001	0.00011

Assume the size of the alphabet is $d > 2$, and $(p_i)_{i=1}^d$ are the probabilities of occurrences of the letters. Let $(X_i, Y_i)_{i=1}^d$ denote the number of occurrences of some letter (in the alphabet) in the two sequences. Then

$$D_2 = X_1Y_1 + X_2Y_2 + \cdots + \left(n - \sum_{i=1}^{d-1} X_i\right)\left(n - \sum_{i=1}^{d-1} Y_i\right),$$

with $\mathbf{E}X_i = np_i$, and $\mathrm{Var}(X_i) = np_i(1 - p_i)$.

When d is large, D_2 is therefore a sum of d identically distributed random variables. If they were independent, then asymptotic normality follows from the usual central limit theorem. Therefore it is natural to conjecture asymptotic normality. Work with G. Reinert using yet another modification of Stein's method, exchangeable pairs coupling, has established this result. For small d and large k the quantity 4^k is large, and D_2 is the sum of 4^k identically distributed terms, and again asymptotic normality is a natural conjecture. We have also made progress on this result.

6. Numerical Experiments

We will now discuss simulation results that support the derived asymptotics. Our simulations were conducted on randomly generated sequences of a fixed length over the alphabet $(\{a, c, g, t\})$ where $d = 4$. The sequences are of independent letters with two distributions of relevance to biological sequence analysis, the uniform distribution ($P(a) = P(c) = P(g) = P(t) = \frac{1}{4}$) and a "G+C-rich" nonuniform distribution ($P(a) = P(t) = \frac{1}{6}$, $P(g) = P(c) = \frac{1}{3}$).

For each length n we generated $2 \times 2{,}500$ sequences and computed the D_2 statistic for each k using our own software. The distribution of the 2,500 scores were then compared to both a compound Poisson process and a normal distribution, using the Kolmogorov–Smirnov test (26) to obtain a p value. As the compound Poisson process does not have an exact computationally efficient distribution, we compared the D_2 simulation results to 2,500 samples from a compound Poisson simulation with appropriate parameters for the given values of n, k, and p_2, using the two-sample Kolmogorov–Smirnov test. When the distributions match, the p values will be distributed on (0, 1) uniformly. When the fit is poor the p values will be near 0.

The results corresponding to the two sequence generation models and the two tests are in Tables 1, 2, 3, and 4. Note that n is taken on a logarithmic scale; in the first columns $n = 100$ and in the last columns $n = 128 \times 100$.

For the normal approximations in the nonuniform case, we expect substantial p values for $k < \frac{1}{6}\log_{1/p_2}(2^x \times 10^2) \sim x/10 + 0.6$. For example, for $x = 2$, we expect normality for $k < 0.8$. While this is observed in our numerical experiments, the results suggest that our bounds are by no means tight.

For normality with uniform sequences, for the $k = 1$ case, D_2 should fail to be normal. To explain the ray of nonzero values, we believe we need to have 4^k large enough but also need $4^k < n = 2^x \times 100$ or $k < x/2 + 3.32$.

For the compound Poisson approximations in the nonuniform case, the derivations suggest approximate compound Poisson when $k > 2\log_{1/p_2}(2^x \times 10^2) \sim 1.2x + 7.2$. In the uniform case, the derivations suggest approximate compound Poisson when $k > \frac{3}{2}\log_{1/p_2}(2^x \times 10^2) \sim 0.75x + 4.95$. These bounds appear to be quite tight.

We did not expect to see the compound Poisson approximation work in the $4^k \sim n$ or the $4^k \ll n$ regimes in the uniform case. In hindsight, we observe the compound Poisson distribution we have chosen approaches a normal distribution (with the

Table 3. Kolmogorov–Smirnov *p* values for nonuniform D_2 compared with compound Poisson

k/n	$2^0 \times 10^2$	$2^1 \times 10^2$	$2^2 \times 10^2$	$2^3 \times 10^2$	$2^4 \times 10^2$	$2^5 \times 10^2$	$2^6 \times 10^2$	$2^7 \times 10^2$
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
3	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
4	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
5	0.15181	0.00010	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
6	0.01670	0.06138	0.00067	0.00000	0.00000	0.00000	0.00000	0.00000
7	0.08230	0.24738	0.12469	0.00075	0.00000	0.00000	0.00000	0.00000
8	0.78766	0.67140	0.04178	0.14229	0.00667	0.00001	0.00000	0.00000
9	0.24738	0.36250	0.13325	0.67140	0.20728	0.01066	0.00179	0.00000
10	0.50706	0.57613	0.06613	0.52972	0.02357	0.95655	0.14229	0.03027

Table 4. Kolmogorov–Smirnov *p* values for uniform D_2 compared with compound Poisson

k/n	$2^0 \times 10^2$	$2^1 \times 10^2$	$2^2 \times 10^2$	$2^3 \times 10^2$	$2^4 \times 10^2$	$2^5 \times 10^2$	$2^6 \times 10^2$	$2^7 \times 10^2$
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
2	0.00059	0.00029	0.00006	0.00409	0.00003	0.00023	0.00549	0.00075
3	0.07657	0.05693	0.02787	0.00605	0.29296	0.02787	0.11658	0.15181
4	0.30940	0.69525	0.03027	0.42112	0.59975	0.10167	0.46307	0.88737
5	0.18346	0.78766	0.11658	0.67140	0.74228	0.29296	0.96602	0.18346
6	0.00104	0.64747	0.13325	0.19509	0.04885	0.69525	0.90400	0.26195
7	0.46307	0.23342	0.46307	0.04885	0.08230	0.48483	0.71892	0.22006
8	0.48483	0.27715	0.03562	0.76524	0.30940	0.91930	0.96602	0.24738
9	0.83039	0.48483	0.19509	0.00199	0.11658	0.74228	0.34418	0.08230
10	0.14229	0.23342	0.78766	0.62356	0.08230	0.07657	0.22006	0.16183

correct mean and variance) when $4^k \ll n$, and this should have been something to be expected, though we still have no theoretical justification for its quality when $4^k \sim n$.

In the nonuniform case, the compound Poisson process also approaches a normal when $4^k \ll n$. However, the variance of this normal distribution does not match the variance of D_2, which explains the lack of fit.

7. Extreme Value Statistics

We have begun to address the issue of the distribution of the statistic D_2. However, when doing database searches, we perform m different comparisons, and we wish to know the p value of the best score of that collection of comparisons. Therefore the distribution of the maximum of m independent values is a desirable feature of a similarity statistic. That is, one often wishes to assign a probability to the maximum of a set of scores being larger than a given value. This can be the maximum score of a single query against a large database of candidates or the maximum of scores coming from "sliding" a query along a long genome sequence. The last important case is even more difficult because of dependencies. For an illustrative example of the technical challenge see the work on a profile score distribution (27).

For a random variable X, approaching a standard normal, the asymptotic extreme value distribution is

$$P(M \leq x) = G(x) = \exp(-e^{-x})$$

with the standardized variable $M = \sqrt{2\log(m)}N^{(m)} + 2\log(m) + \frac{1}{2}\log(\log(m)) + \frac{1}{2}\log(4\pi)$, for the maximum value, $N^{(m)}$, of m independent standard normal random variables.

Using this for intuition, we can explore the fit from treating our approximately normal random variables for D_2 as if they were independent normals. This was done in ref. 27 for profiles.

In the figures for this section we make a plot of $-\log(-\log(G(x)))$ for the cumulative distribution function of 2,500 samples of the maximum of 100 D_2 scores on both uniformly distributed and nonuniform sequences for various (k, n) values. These plots should approach the line $y = x$ (for $x > 0$) in the asymptotic limit. We chose four values for k and n that were just within the experimentally determined normal regime, $(k, n) = (6, 2^4 \times 10^2)$, $(7, 2^5 \times 10^2)$, $(8, 2^6 \times 10^2)$, $(9, 2^7 \times 10^2)$. We have also plotted four *null plots* that are the result of taking 2,500 samples of the maximum of 100 random variables from a true normal distribution for comparison (Fig. 1).

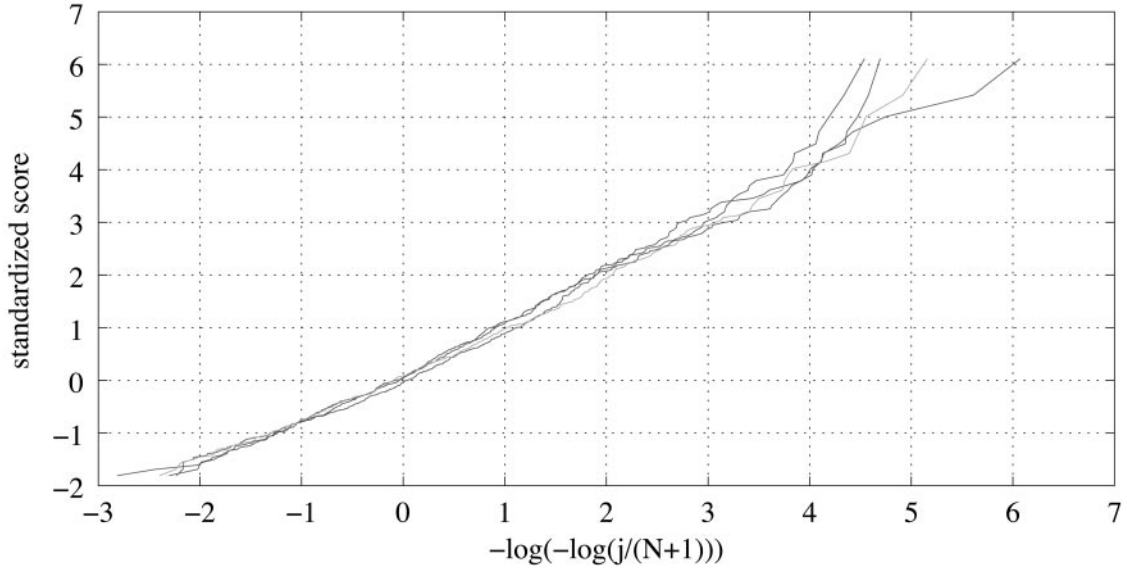

Fig 1. The extreme value plot for true normal data.

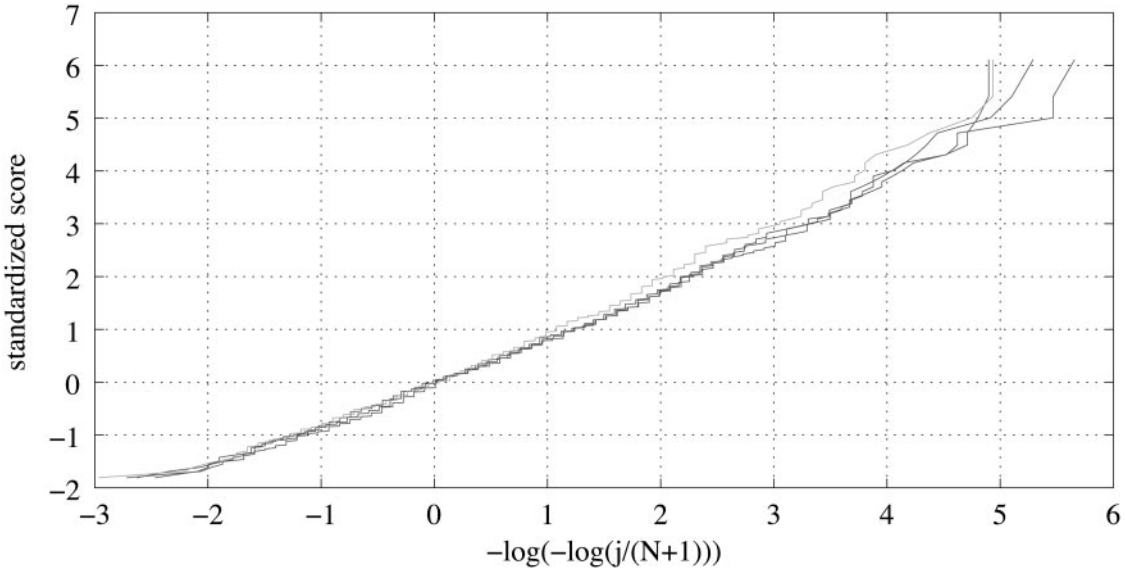

Fig 2. The extreme value plot for D_2 on uniform sequences.

We see that Fig. 2 shows the uniform D_2 plot converging very well to the extreme value distribution. We see that Fig. 3 shows poor convergence even though the (k, n) values are well within the normal range, as shown by Table 1. This is illustrative of the technical challenges of establishing rigorous results for these cases. In addition, we point out the famously slow convergence of the maximum of random variables to their extreme value limiting distribution.

8. Conclusion

We have established a model for the distribution of a word composition (dis)similarity statistic D_2 for the case where the underlying sequences are nonuniformly distributed, and we have numerically investigated this statistic on both uniform and nonuniform sequences. The statistical limits of (compound) Poisson when word matches are rare and of normal when word matches are common are observed. However, there are "small-word" regimes where D_2 is neither, and thus convergence is both a matter of word length and match probability.

In *Section 6*, we found that for a nucleotide sequence of EST length (~500), and with the typically selected $k = 6$ two-codon word-size, the compound Poisson is a better fit than the normal *unless the letter distribution is uniform or close to uniform* (where it is shown, by experiment, to be equivalent).

In *Section 5*, we saw that D_2, in the nonuniform asymptotics for $k = 1$ and binary alphabet, standardizes to two independent normals, each of which is dependent only on one sequence and the expected distribution of words in the model. To contribute asymptotically to the sum D_2, the product of the standardized count statistics must grow at least as fast as $\sqrt{n}$. Otherwise "unusual" k-word coincidences between the sequences will be masked. These observations hold for larger k and for nonbinary

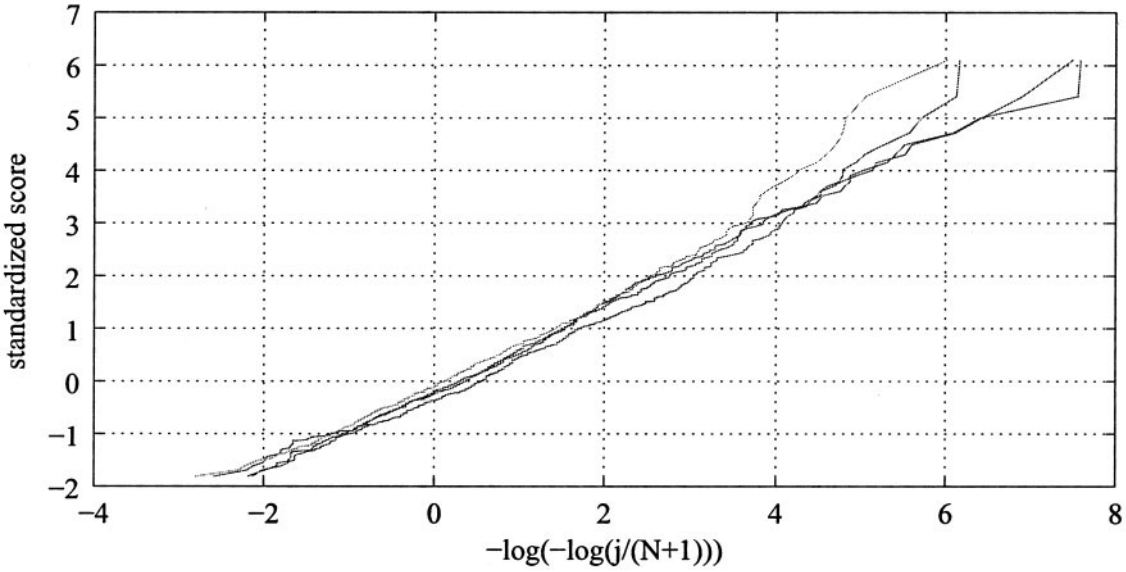

Fig 3. The extreme value plot for D_2 on nonuniform sequences.

alphabets, within certain restrictions on the growth of n. Thus, anyone hoping to simply use inner products of word compositions is likely going to be measuring the sum of the departure of each sequence from the background, which seems to miss the point of sequence comparison. This mathematics suggests that, asymptotically, one could gain as much insight into pairwise similarity of nonuniform sequences by pairwise comparing some appropriate indices of background departure, computed per each sequence.

This result raises the question of which of various compositional statistics could have a tractable distribution and an asymptotic relevance to pairwise comparison. One possibility we have yet to explore is that of the D_2 variant used originally by Blaisdell in ref. 6, where distance replaces the inner product in the comparison of composition vectors.

1. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147,** 195–197.
2. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448.
3. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
4. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
5. Waterman, M. S. (1995) *Introduction to Computational Biology* (Chapman & Hall, New York).
6. Blaisdell, B. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 5155–5159.
7. Torney, D. C., Burke, C., Davison, D. B. & Sirotkin, K. M. (1990) in *Computers and DNA*, SFI Studies in the Sciences of Complexity, eds. Bell, G. & Marr, T. (Addison-Wesley, Reading, MA), Vol. 7, pp. 109–125.
8. Christoffels, A., Gelder, A., Greyling, G., Miller, R., Hide, T. & Hide, W. (2001) *Nucleic Acids Res.* **29,** 234–238.
9. Carpenter, J. E., Christoffels, A., Weinbach, Y. & Hide, W. A. (2002) *J. Comput. Chem.* **23,** 1–3.
10. Hide, W., Burke, J. & Davison, D. B. (1994) *J. Comput. Biol.* **1,** 199–215.
11. Burke, J., Davison, D. B. & Hide, W. (1999) *Genome Res.* **9,** 1135–1142.
12. Ji, H., Zhou, Q., Wen, F., Xia, H., Lu, X. & Li, Y. (2001) *Nucleic Acids Res.* **29,** 260–263.
13. Mironov, A. & Alexandrov, N. (1988) *Nucleic Acids Res.* **16,** 5169–5173.
14. Wu, T., Burke, J. & Davison, D. (1997) *Biometrics* **53,** 1431–1439.
15. Wu, T., Hsieh, Y. & Li, L. (2001) *Biometrics* **57,** 441–448.
16. Stein, C. (1972) *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (Univ. California Press, Berkeley), Vol. 2, pp. 583–602.
17. Chen, L. H. Y. (1975) *Ann. Probab.* **3,** 534–545.
18. Schbath, S. (1995) *ESAIM: Probab. Stat.* **1,** 1–16.
19. Reinert, G. & Schbath, S. (1998) *J. Comput. Biol.* **5,** 223–253.
20. Borbour, A. & Chryssaphinou, O. (2001) *Ann. Probab.* **11,** 964–1002.
21. Stein, C. (1986) *Approximate Computation of Expectations* (Inst. Mathematical Statistics, Hayward, CA).
22. Dembo, A. & Rinott, Y. (1994) *IMA Vol. Math. Appl.* **76,** 25–44.
23. Rinott, Y. & Rotar, V. (1996) *J. Multivariate Anal.* **56,** 333–350.
24. Rinott, Y. & Rotar, V. (2000) *Decisions in Economics and Finance* **23,** 15–29.
25. Chen, L. H. Y. & Shao, Q.-M. (2002) *Normal Approximation Under Local Dependence*, preprint.
26. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992) *Numerical Recipes in C* (Cambridge Univ. Press, New York), pp. 620–630.
27. Goldstein, L. & Waterman, M. S. (1994) *J. Comput. Biol.* **1,** 93–104.